Neutron and gamma discrimination based on multi-features and KNN-LDA)*

Yuhang Jiang,[1] Tingmeng Ding,[1] Xuanxi Wang,[1] and Xiaofei Jiang[1, †]

[1]*Big Data and Information Engineering, GuiZhou University, Guiyang 550025, China*

In order to improve the n/γ discrimination effect, this paper proposes an intelligent discrimination method based on multi-features and KNN-LDA algorithm. Firstly, this paper establishes 6 feature parameters covering time and frequency domains according to the PSD principle; secondly, an automatic feature extraction system is designed, which can intelligently extract the optimal distributions of multi-features from the pulse data. Then, the feature criterion is constructed and calculated based on the feature distributions, so as to divide reliable data from the feature data as the training set of the model, and the rest as the test set. Finally, based on the training set, using the regression optimization and dimensionality reduction of the KNN-LDA model, the test set is input into the model for further classification to achieve the n/γ discrimination of all pulses. According to the experimental results, the FOM value of the KNN-LDA model reaches 3.07 in the test set of high-energy domain ($\geq$40 keV), which is 245% higher than that of the CCM; in the test set of low-energy domain ($\leq$40 keV), the FOM value of the KNN-LDA model reaches 2.64, which is 355% higher than that of the CCM. The experimental results show that the discrimination effect of this method is excellent, especially the discrimination of low-energy-domain is greatly improved, which provides a new idea for introducing supervised learning model to solve the difficult problem of low-energy-domain discrimination.

Keywords: N/γ discrimination, Multi-features, Linear discriminant analysis(LDA), K-nearest neighbor(KNN), Automatic feature extraction

I. INTRODUCTION

The discrimination between neutrons and γ-rays is a key issue in neutron detection research [1–3]. The discrimination method based on the PSD (Pulse shape discrimination) idea has been the mainstream method [4–6], and in recent years, with the rapid development of machine learning, the intelligent discrimination method based on machine learning has also become a hot research direction in this field [7–9]. The traditional PSD-based discrimination methods usually use a single feature to visualize the pulse shape difference [10–12], and use the distribution of the feature to characterize the differentiation effect between neutrons and γ-rays to achieve n/γ discrimination. However, when the pulse shape difference between neutrons and γ-rays is small, especially when the difference expressed in this feature parameter is small, the discrimination ability of this feature is greatly reduced, which always occurs in the low-energy domain [13–15], and is also the main reason why it is harder to discriminate the pulse data in the low-energy domain. Intelligent discrimination methods based on machine learning have been used as an effective way to enhance the discrimination effect of low-energy-domain data, among which supervised learning algorithms, which have the advantages of strong interpretability, high prediction accuracy, and wide applicability, have been applied many times in this field [16–18]. However, supervised learning algorithms often need reliable learning samples to train the model, but for the n/γ discrimination field, there is no standard database of neutron and γ-ray pulses to provide samples for the model to train, which makes the portability of the supervised learning model greatly weakened, and also significantly increases the difficulty of the model applied to the field discrimination.

Aiming at the pulse data that are difficult to discriminate, in order to effectively improve the n/γ discrimination effect and increase the robustness of the model, this paper proposes an intelligent discrimination method based on multi-features and KNN-LDA. First, this paper establishes six waveform feature parameters to jointly characterize the pulse differences from multiple dimensions, and designs an intelligent feature extraction algorithm to automatically obtain the optimal distribution of each feature. Secondly, the distribution of each feature is synthesized to construct a feature criterion, which is used to classify the part of feature data that is reliable and easy to be discriminated, so as to complete the preliminary classification, and this part of the feature data is turned into the training set, and the remaining data is used as the test set. Then, the KNN-LDA model is trained with the training set, and regression optimization and dimensionality reduction classification are carried out on the remaining test set, and n/γ discrimination of all pulses is finally realized.

II. METHODS

A. Selected multi-features

In order to improve the discrimination accuracy, this study proposes to construct feature parameters covering both time and frequency domains to participate in n/γ discrimination, but from the viewpoint of the discrimination model, the computational complexity of the feature parameters must be taken into account. If the construction and computation of each feature parameter is cumbersome, then accordingly the application of the whole multi-feature parameters will face difficulties in feature extraction, which will undoubtedly make the

* Supported by the National Science Fund for Distinguished Young Scholars of China (No.12205062) and Network Communication Signal Detection System(No.1502195N)

† Corresponding author, 514651931@qq.com

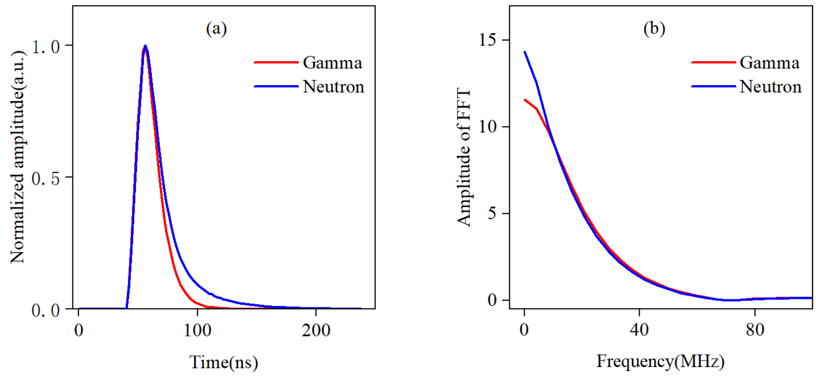

Fig. 1. Pulse shape difference: (a) time domain; (b) frequency domain.

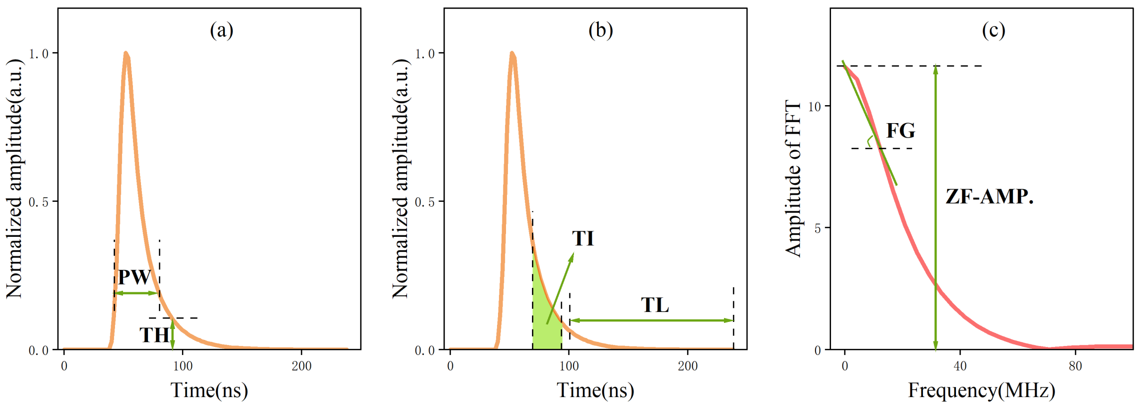

Fig. 2. Selected features: (a) pulse width(PW) and tail height(TH); (b) tail integral(TI) and tail length(TL); (c) frequency gradient(FG) and zero frequency amplitude(ZF-AMP.).

study lose sight of the problem. Thus, the feature parameters we select should be able to express the pulse differences in different dimensions, but also to meet the needs of simple construction and easy extraction, so that it is possible to carry out efficient multi-features extraction, which is the premise of the application of multidimensional feature parameters to n/γ discrimination.

Normalized neutron and γ-ray pulses usually differ in the falling edge segment [19–21], as shown in Fig. 1. a. It is intuitively obvious that this difference can be concretized in the transverse direction, in the longitudinal direction, and in the integral. Accordingly, we construct two features of pulse width (PW) and tail length (TL) in the transverse direction, where pulse width does not narrowly refer to the width of the pulse at 50% amplitude, but refers to the width of the pulse and a certain amplitude level from the intersection point of the rising edge to the intersection point of the falling edge, and similarly the tail length, i.e., refers to the width of a certain sample point at the falling edge to the last sample point of the pulse; and we construct a feature of the tail height (TH) in the longitudinal direction, which is the amplitude of the pulse at a certain point along the falling edge; and the tail integral (TI) feature is also constructed, i.e., the integral value of a certain part of the falling edge. The four time-domain features are shown in Fig. 2. a and Fig. 2. b. The normalized neutron and γ-ray pulses are obviously different in the low-frequency range of the amplitude spectrum after the FFT [22, 23], as shown in Fig. 1. b. Drawing on the PSD features of the traditional frequency-domain discrimination method, we constructed zero-frequency amplitude (ZF-AMP.) and frequency gradient (FG), which were utilized by the articles[60,61] to effectively complete the n-γ discrimination, which demonstrated that they have a good discrimination ability, as shown in Fig. 2. c.

This study combines the traditional discrimination methods, selects the feature parameters from multiple viewpoints in the time-frequency domain to comprehensively characterize the difference information contained in the pulse waveforms, and applies the constituted multidimensional feature parameters to the n-γ discrimination study, with a view to obtaining better discrimination results, especially in the case of smaller differences in the pulse waveforms, to obtain better discrimination effects and more reliable classification results.

B. Automatic feature extraction

In PSD-based discrimination methods [24–26], the distribution of different PSD factors is generally obtained by continuously adjusting the feature measurement parameters, from which the optimal distribution is determined, and the valley between the two distribution peaks is used as the discrimination threshold to realize n/γ discrimination. Among them, the process of finding the optimal distribution of features needs to be manually adjusted, and the optimal parameters under different experimental scenarios are different [27–29], so this process is sometimes cumbersome and repetitive, and because of human judgment, the best results may not be obtained in the end. Therefore, it is crucial to realize the automation of feature extraction [30].

The key to the automation of feature extraction is to find and output the feature distribution that offers the best separation, so it is necessary to compare the separation degree of the feature distribution obtained from different measurement parameters. In the traditional methods, FOM (Figure Of Merit) is used to evaluate the n/γ discrimination effect [31], i.e., it is used to express the degree of separation between the neutron peak and γ-ray peak, and in order to make a distinction, DOS (Degree Of Separation) is used in this paper, which is defined as shown in Fig. 3, in which d is the spacing of the two peaks, and $FWHM_n$ and $FWHM_\gamma$ are the full width at half maximum of the neutron peak and the γ-ray peak, respectively.

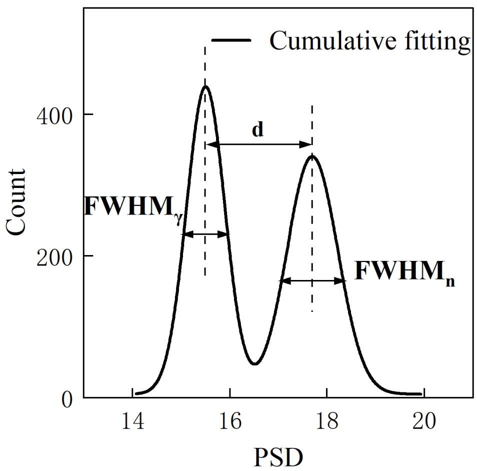

Fig. 3. The sketch map of DOS.

The DOS value can be calculated by Eq. (1), the larger the DOS value, the greater the degree of separation. By calculating and comparing the DOS values of the feature distribution measured with different parameters, from which we can find the optimal feature distribution we need, thus realizing the automated extraction of feature.

$$DOS = \frac{d}{FWHM_n + FWHM_\gamma} \tag{1}$$

The automatic extraction process of a single feature is shown in Fig. 4: firstly, based on the input preprocessed pulse and the feature to be measured, the range of extraction parameters for the feature is confirmed; secondly, the loop structure is entered, and a set of measurement parameters is inputted into each loop, and the program measures the features of all the pulses according to the parameters, thus obtaining a feature distribution; then, the DOS values of the neutron peak and the γ-ray peak are calculated, which is compared with the DOS value stored in the previous loop in the shift register, leaving the larger value with its corresponding feature measurement value; finally, when the loop ends, the largest value and the corresponding feature measurement value are outputted. Then, calculate the DOS value of the neutron peak and the γ peak, compare it with the DOS value stored in the previous cycle in the shift register, and leave the larger value with its corresponding feature measurement value to enter the next cycle; finally, when the cycle is finished, output the largest DOS value and the corresponding feature value. That is to say, the automatic extraction of feature is accomplished while ensuring the best degree of separation.

In the above process, the most critical thing is how the program autonomously calculates the DOS. Due to the indeterminate shape of the feature distribution curves and the inability to directly obtain the calculation parameters of the DOS, we uniformly perform Gaussian fitting on the distribution curves and use the fitting parameters to directly calculate the DOS, as in Eq. (2), where μ_n, μ_γ, and σ_n, σ_γ are the mean and standard deviation of the peaks of the neutron fitting and the γ fitting, respectively. The mean and standard deviation of the fits can be obtained directly after performing Gaussian fitting.

$$DOS = \frac{|\mu_n - \mu_\gamma|}{2.355 \times (\sigma_n + \sigma_\gamma)} \tag{2}$$

However, for the feature distribution curves, the program cannot perform a Gaussian fit of the bimodal peaks by itself, so we need to switch the idea to split the distribution bimodal peaks from the troughs between them, thus forming two single peaks in order to perform a Gaussian fit of the single peaks automatically at the same time. In this way, the problem converts into how to find the correctly fitting trough.

Because the feature distribution curves are affected by many factors, such as the size of the pulse dataset, the feature category, the measurement location of the same feature, etc., the shape of the distribution curve is not necessarily a regular bimodal shape, and it is more likely to have a variety of burrs and fluctuations, which makes it difficult to determine the location of the target trough, beacuse of the presence of a very large number of meaningless troughs during the trough searching. Therefore, before seeking the trough, this paper preforms three times spline fitting (the balance parameter is set to 0.9), so that the distribution curve tends to be smooth and will not be deformed. Then according to the position of the double peaks stuck in the position of the target valley, using the determined valley to split the original distribution curve before fitting, and then at the same time for Gaussian fitting, and then calculate the DOS value. In this way, the program can autonomously find the best feature distribution,

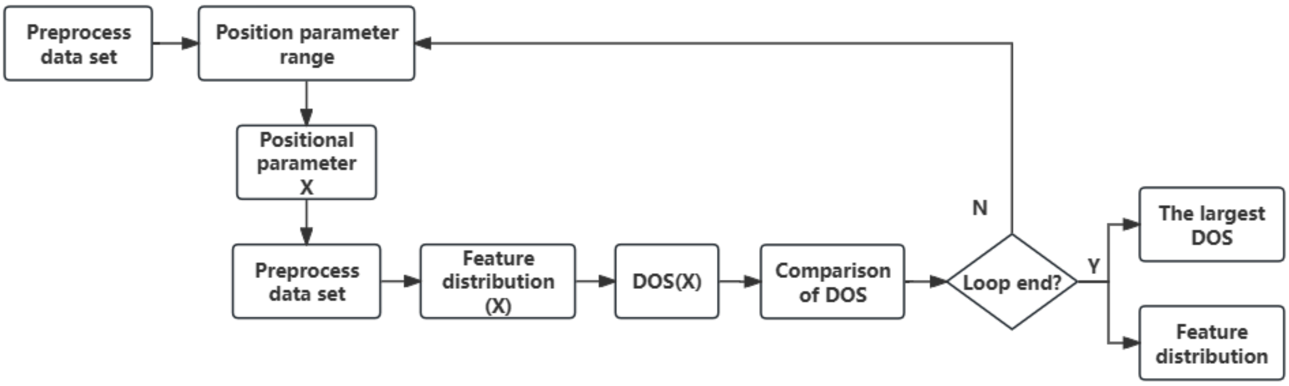

Fig. 4. The flow chart of feature extraction.

thus optimizing the process of manual parameter tuning.

C. Feature criterion

After obtaining the feature data of the pulse, each feature will be classified by the feature distribution. In this paper, Comprehensive Voting Factor (CVF) and Comprehensive Location Factor (CLF) are constructed to comprehensively evaluate the performance of each pulse in the classification of multi-features, and based on them, reliable data sets are separated from the feature data and used as the training samples, and the remaining data as relatively more difficult to distinguish test samples, and then use the trained KNN-LDA to further discriminate the test set.

The initial discrimination through the feature distributions from the automatic feature extraction (discrimination for neutron marked as 1, for γ-ray marked as 0), to get 6 discrimination results data set, from which we can construct a matrix of 0 and 1 composed of $M \times 6$ to represent these data results, as in Eq. (3). The computed weights W of CVF are calculated by Eq. (4). We can obtain a matrix $B_{6\times1}$ as in Eq. (5), and the CVF can be calculated as in Eq. (6).

$$A_{M\times6} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \tag{3}$$

$$W_{feature} = \frac{DOS_{feature}}{\Sigma DOS} \tag{4}$$

$$B_{6\times1} = (W_{PW}, W_{TH}, \ldots, W_{ZF-AMP.}, W_{FG})^T \tag{5}$$

$$CV_{M\times1} = A_{M\times6} \times B_{6\times1} \tag{6}$$

The values $CVF(i)$ in the $CV^T_{M\times1}$ matrix all range from 0 to 1. We use Eq. (7) to determine the category of the ith pulse, i.e., we realize the comprehensive voting result with multi-features.

$$\begin{cases} \gamma & 0 \le CVF(i) < 0.5 \\ n & 0.5 \le CVF(i) \le 1 \end{cases}, i = 1, 2, \ldots, M \tag{7}$$

Obviously, if $CVF(i) = 0$, all six features classify the pulse as a γ-ray pulse; if $CVF(i) = 1$, all six features classify it as a neutron pulse. For pulses with a CVF of 0 and 1, we consider their corresponding discrimination results to be very reliable and can be used to construct the training set. The closer the CVF is to 0.5, the more difference occurs in multi-features discrimination, and the less reliable the obtained results are.

The CVF integrates the category judgment of each feature for each pulse, and for the feature distribution output from the automatic feature extraction system, specifically for each feature of each pulse in the corresponding feature distribution, the pulse is likely to have different performances in each feature, that is to say, the pulse may be far away from the trough in this feature distribution, and immediately adjacent to the trough in another feature distribution, so sometimes a single feature may not be able to fully represent the differences in pulse shapes. On the other hand, the higher the quality of the training samples for the supervised model, the better the model performance will be. Therefore, we construct the CLF to further optimize the reliable data selected according to the CVF from the relative positions of the feature distributions, and thus obtain a high-quality training set.

As Fig. 5 shows the distribution curve of a typical PSD feature histogram, assuming that the left peak is the γ-peak and the right peak is the neutron peak, the value of this PSD feature for the ith pulse is c(i), and the trough value is v. Based on this, we define the relative position of each pulse on this feature in terms of Eq. (8):

$$L(i)_{feature} \begin{cases} \frac{c(i)-v}{a}, c(i) < v \\ \frac{c(i)-v}{b}, c(i) \ge v \end{cases}, i = 1, 2, \ldots, M \tag{8}$$

where $L(i)_{feature}$ refers to the relative position of the ith pulse on the feature; a and b are the distances from the gamma

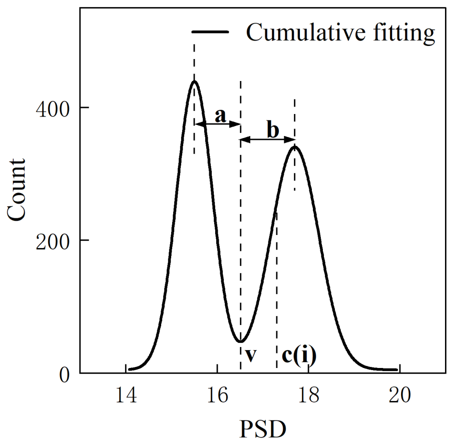

Fig. 5. Schematic diagram of the relative position.

and neutron peaks to the trough, respectively; and M represents the number of mixed pulses.

Obviously, the relative position of the pulse with the feature value located at the γ peak is negative, the relative position of the pulse located at the neutron peak is non-negative, and the relative position of the pulse located at the center of the two peaks is exactly ± 1, i.e., the larger the absolute value of the relative position and the further away from the zero point, the better the separation is at that feature. The distribution of the relative positions of a single feature is fundamentally similar to the shape of the original distribution, which gives the conditions to synthesize the feature distributions of each pulse, and we define the CLF directly in Eq. (9):

$$CLF(i) = \frac{L(i)_{feature}}{6} \tag{9}$$

It can be seen that the larger the absolute value of CLF is, the farther the relative position of the pulse is from the classification boundary point, the better its combined performance on the feature distributions is, and the higher the corresponding classification reliability is.

To summarize, we use CVF and CLF together as constraints to form the feature criterion as in Eq. (10). CVF is constrained in terms of multi-features category reliability and CLF is constrained in terms of multi-features positional reliability, which selects a high-quality feature training set.

$$\begin{cases} CVF(i) = 0 \cup CVF(i) = 1 \\ |CLF(i)| > 0.5 \end{cases} , i = 1, 2, \ldots, M \tag{10}$$

D. KNN-LDA

K-Nearest Neighbor (KNN) is a commonly used supervised learning method [32], and its working mechanism is very simple: given a test sample, find out the closest K samples in the training set based on a certain distance metric, and then make a prediction based on the information of these K "neighbors". Obviously, KNN is a kind of inert learning model, its training phase is only to save the training samples, and then calculate the corresponding distance after receiving the test samples, and because each test sample has to calculate the distance to all the training samples, which makes its testing time overhead is large. Therefore, in this paper, the six extracted features are used as inputs to the KNN to reduce the computation of the algorithm while ensuring that the impulse difference information is fully retained. In this paper, the distance between test samples and training samples is calculated based on the Euclidean distance as in Eq. (11). where $D(X, Y)$ is the distance between sample X and sample Y, m is the number of features contained in the samples, and x_i and y_i are the corresponding ith feature values in the two samples, respectively.

$$D(X, Y) = (\Sigma_1^m (|x_i - y_i|)^2)^{1/2} \tag{11}$$

After distance calculation and comparison, the K training samples closest to the test samples are obtained, and the average of these K samples is used as the regression result using the "averaging method". In this way, based on the training set of feature data, the test set can be regressed by KNN, and the multi-dimensional characteristics of feature data can support the reasonableness of the regression features used for classification, so that the regression features can be used as the input of LDA for further dimensionality reduction and classification.

Linear discriminant analysis(LDA) [33] is a classical linear learning method, which was first proposed by Fisher in 1936 for binary classification problems. Moreover, the projection process of LDA has the effect of dimensionality reduction and retains the category information, so LDA is also a classical supervised dimensionality reduction technique. The idea of LDA [34] is to project a given set of training samples onto a straight line, so that the projection points of samples of the same class are as close as possible, and the projection points of samples of different classes are as far away from each other as possible; when a new sample is classified, it is projected onto this straight line, and the class of the sample is determined according to the position of its projection point, as shown in Fig. 6.

The pulse data can be replaced by the matrix of Eq. (12), which represents a total of m digitized pulses of neutrons mixed with γ-rays, each discretized into l sampling points. Specify that $x_n = (x_{n1}, x_{n2}, \ldots, x_{nl})$ denotes the neutron pulse, and $x_\gamma = (x_{\gamma 1}, x_{\gamma 2}, \ldots, x_{\gamma l})$ denotes the γ-ray pulse.

$$X_{M \times l} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{ml} \end{pmatrix} \tag{12}$$

Let the mean vectors and covariance matrices of x_n, x_γ be μ_n, μ_γ and σ_n, σ_γ respectively, if the pulse data are projected onto a straight line w, then the projections of the centers of the two classes of samples onto the straight line w will be $w^T \mu_n$ and $w^T \mu_\gamma$, and the covariances of the two classes of samples will be $w^T \Sigma_n w$ and $w^T \Sigma_\gamma w$ respectively. Our goal is to

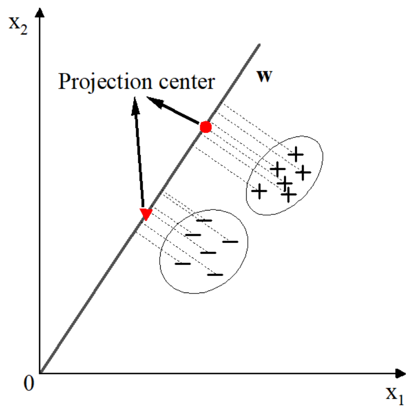

Fig. 6. The sketch map of LDA.

maximize the inter class distance between two types of projection points and minimize their intra class distance, i.e., to make the center distance $\|w^T\mu_n - w^T\mu_\gamma\|_2^2$ as large as possible, and the sum of the covariances, $w^T\Sigma_n w + w^T\Sigma_\gamma w$ as small as possible, if we consider the two at the same time, we can get the goal of maximization:

$$\begin{aligned} J &= \frac{\|w^T\mu_n - w^T\mu_\gamma\|_2^2}{w^T\Sigma_n w + w^T\Sigma_\gamma w} \\ &= \frac{w^T(\mu_n - \mu_\gamma)(\mu_n - \mu_\gamma)^T w}{w^T(\Sigma_n + \Sigma_\gamma)w} \end{aligned} \tag{13}$$

Let the within-class scatter matrix be $S_w = \Sigma_n + \Sigma_\gamma$, the between-class scatter matrix be $S_b = (\mu_n - \mu_\gamma)(\mu_n - \mu_\gamma)^T$, which is brought into Eq. (13), we can obtain the Eq. (14):

$$J = \frac{w^T S_b w}{w^T S_w w} \tag{14}$$

By derivation, it can be calculated that when the line w satisfies Eq. (15), J obtains a maximum value.

$$w = S_w^{-1}(\mu_n - \mu_\gamma) \tag{15}$$

Wherein, considering the stability of the numerical solution, S_w^{-1} is usually obtained by singular value decomposition of S_w, i.e., $S_w = U\Sigma V^T$, where Σ is a real diagonal matrix whose elements on the diagonal are the singular values of S_w, and after that S_w^{-1} can be obtained from Eq. (16). From this, the optimal projection direction w can be obtained.

$$S_w^{-1} = V\Sigma^{-1}U^T \tag{16}$$

From the above principle analysis, we can see: KNN regression algorithm and LDA classification algorithm principle is simple, easy to calculate, easy to apply, we can make full use of and combine their regression optimization and dimensionality reduction classification function. In summary, the KNN-LDA model is to obtain the KNN regression of the test set through the training set data and get the optimal projection vector of LDA, and then use the optimal projection vector to project the regression features down to a one-dimensional space, and then get the discrimination results of one-dimensional features. In this paper, through the intelligent extraction of multi-features and the construction and computation of feature criterion, high-quality training set can be separated instantly, which greatly facilitates the classification of KNN-LDA.

III. EXPERIMENT

LabVIEW is a graphical development environment, is commonly used to build digital n-γ discrimination platform programming software, which can build a user interactive interface to facilitate the signal display and parameter changes. Moreover, its graphical programming is simple and easy to learn, easy to start, but also embedded in a variety of directly usable data processing modules, greatly reducing the difficulty of programming, and is widely used in industrial automation, test and measurement, signal processing and other fields [35–37]. As shown in the Fig. 7, the flowchart of multi-features based KNN-LDA model discrimination constructed by LabVIEW is shown in this paper.

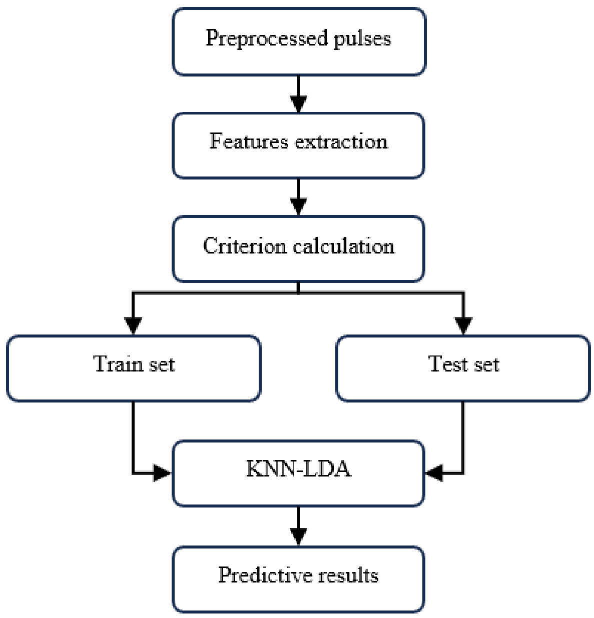

Fig. 7. The flow chart of KNN-LDA.

First, input the preprocessed pulse data, from which the automatic feature extraction algorithm extracts multi-features. Then, calculate the corresponding CVF and CLF based on the output multi-features classification result and distribution location, selecting high-quality training samples from the feature data to complete the initial classification. Finally, deliver the training samples to the KNN-LDA model, where the model is trained to perform regression and dimensionality reduction classification on the remaining fuzzy samples (test

set), and the classification results are output, thus completing the discrimination task.

A. Data acquisition and processing

The experimental data set of this paper contains 40,000 pulse data, which is composed of pulse signals obtained by the EJ-301 detector detecting the ^{241}Am-Be neutron source. The neutron detection experimental environment and data acquisition schematic are shown in Fig. 8. During the experiment, the detector is placed horizontally, the angle between the detector and the radioactive source is 0 degree, and the distance between the neutron source and the detector is set to be 40 cm, and the detected pulse signals are processed by a digitizer at 500 MS/s and then transmitted to a PC for preservation, and each pulse waveform contains 128 sampling points, and the time interval of the two neighboring sampling points is 2 ns. Fig. 9.

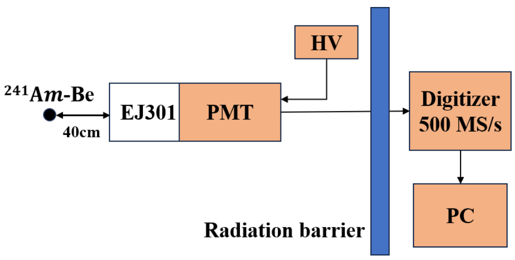

Fig. 8. Schematic diagram of neutron detection experiment.

In the process of converting neutrons and γ-rays into pulse signals through detectors, there will be electronic noise and baseline drift and other factors, so it is necessary to carry out a certain amount of pre-processing of the pulse to try to eliminate these effects and unify the premise of feature extraction. For the pulse dataset in this paper, we mainly carry out the operations of baseline zeroing, flipping, smoothing filtering and normalization.

In this paper, the 30 sample points at the end of the pulse waveform are selected, the mean value of this points is obtained, and then the whole pulse is subtracted from this mean value, that is, a simple baseline zeroing is completed. After that, the amplitude of each sample point of the pulse is added with a minus sign, that is, the pulse is flipped. Then this paper utilizes a smoothing filter to carry out a moving average smoothing filter for the signal at the seven-point scale, which essence is to take a total of 7 points, including 3 points before and after the current point and the point itself, to calculate the mean of their amplitudes, to replace this point; Finally, the pulse is mapped between 0 and 1, which completes the normalization of the pulse. The preprocessing effect is shown in

The pulses with small waveform differences and difficult to discriminate in the pulse dataset are mainly concentrated in the low-energy domain part, in order to explore the ability of the present method to improve the discrimination effect of the pulses in this part, this paper divides the preprocessed pulse dataset into low-energy-domain pulse dataset (containing 21,464 pulses, as dataset 1) and high-energy-domain pulse dataset (containing 18,536 pulses,as dataset 2) with the limit of 40 keV, and carries out the discrimination experiments of the present study, respectively.

B. Features extraction

After inputting the preprocessed pulse into the automatic feature extraction system, the system will automatically obtain the optimal distribution of each feature, and rely on the trough position to automatically perform single-feature discrimination, and the obtained results are shown in Table 1, and the corresponding feature distributions and DOS values are shown in Fig. 10, Fig. 11, and Fig. 12.

As can be seen from Table 1, taking the number of neutrons as an example, in the high-energy domain, the largest difference in the number of particles discriminated by multi-features is between TH and TL, with a difference of 226 particles, accounting for 1.2% of the number of particles; while in the low-energy domain, the largest difference in the number of discriminated particles is between TH and TI, with a difference of 1,747 particles, accounting for 8.1% of the number of particles, which is about 6.8 times of that in the high-energy domain. To put it in another way, if we take the difference in the number of discrimination particles of any two different features to find the mean value, we can get that the mean difference in the high-energy domain is about 91, accounting for 0.49% of the number of particles, while that in the low-energy domain is about 786, accounting for 3.7% of the number of particles, which is about 7.6 times that of the high-energy domain. These analyses illustrate that the discrimination in the low-energy domain has more uncertainties than that in the high-energy domain, i.e., the discrimination in the low-energy domain is more difficult, and it also directly expresses that the reliability of a single feature parameter when performing the discrimination task in the low-energy domain decreases dramatically compared with that in the high-energy domain. Thus, the comprehensive use of multidimensional feature parameters for n/γ discrimination is undoubtedly a direct and effective way to improve the reliability of discrimination in the low-energy domain.

As can be seen from the feature distributions and Fig. 12, the DOS of each feature distribution of the time-domain part obtained in the low-energy domain is significantly less than that in the high-energy domain, and the bimodal separation of the feature distributions in the high-energy domain is much better, which clearly indicates that the high-energy-domain data are more easily discriminated. Moreover, the TI performs poorly in the low-energy domain but well in the high-energy domain, during which the DOS value changes the most, which suggests that it may be more suitable for the high-energy domain discrimination task; on the contrary, the two frequency-domain features have good DOS performance in both the low-energy domain and the high-energy domain, which suggests that the frequency-domain discrepancy combines with a better discrimination ability and a stronger sta-

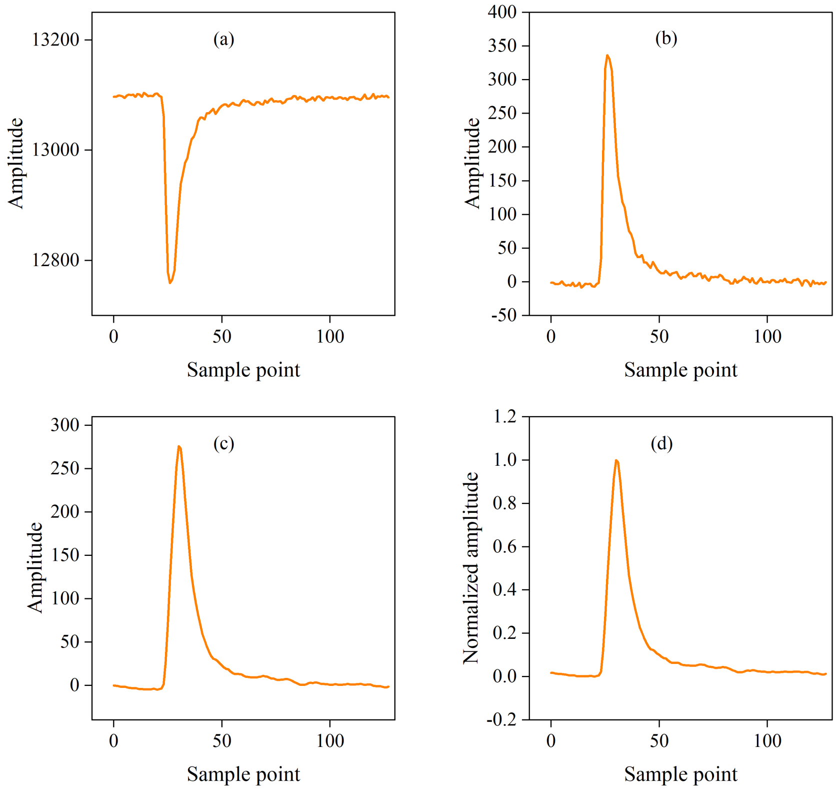

Fig. 9. Signal preprocessing:(a)Original signal;(b) Flipped and baseline restored;(c) 7 points moving smooth;(d) Normalized.

Table 1. The discrimination result of multi-features.

Dataset	Category	PW	TH	TI	TL	ZF-AMP.	FG
1	Neutron	9728	9896	9790	9670	9716	9739
	Gamma	8808	8640	8746	8866	8820	8797
2	Neutron	8222	7198	8945	7206	7888	7883
	Gamma	13242	14266	12519	14258	13576	13581

bility. Therefore, it is an intuitive and effective way to discriminate impulse data with multi-features covering the time-frequency domain and thus improve the discrimination results.

When the multi-features are output, according to the classification results and distribution location of each feature, the program calculates the CVF and CLF of each pulse accordingly, and the resulting distribution histogram is shown in Fig. 13 and Fig. 14.

Fig. 13 and Fig. 14 show the comprehensive classification results of multi-features in terms of categories and distributions, respectively. From the CVF histogram, it can be seen that the pulses with CVF values of 0 and 1 are the main components of the pulse dataset, which represent the same classification voting results of multi-features, whereas the values between 0 and 1 represent the contradiction of multi-features classification voting results, and the closer the CVF value is to 0.5 the deeper the contradiction is. Compared with the high-energy domain, the distribution between 0 and 1 in the low-energy domain plot is obviously piled up, which can be calculated from the data that this part of the pulse in the low-energy domain dataset accounts for 12.23%, and in the high-energy domain dataset accounts for only 2.42%, which is about 10% less than that in the high-energy domain, which obviously verifies the objective situation that it is more difficult to classify the pulse data in the low-energy domain than that in the high-energy domain. And from the CLF histogram, it forms an obvious double-peak shape with zero as the trough and ± 1 as the peak, and the pileup of the low-energy domain at zero is also more serious, and the double-peak separation effect of the high-energy domain is obviously better, which also more intuitively shows that the difference of waveforms of pulses in the low-energy domain is even more minute.

In order to improve the discrimination effect on the harder-to-discriminate impulse data, we introduce the supervised discrimination model of KNN-LDA, and use the feature criterion composed of CVF and CLF to instantly select high-quality training samples for the model. According to Section 2.3, in this paper, a total of 18,071 sets of feature data (including 6,396 sets of neutron feature data and 11,675 sets of γ-ray

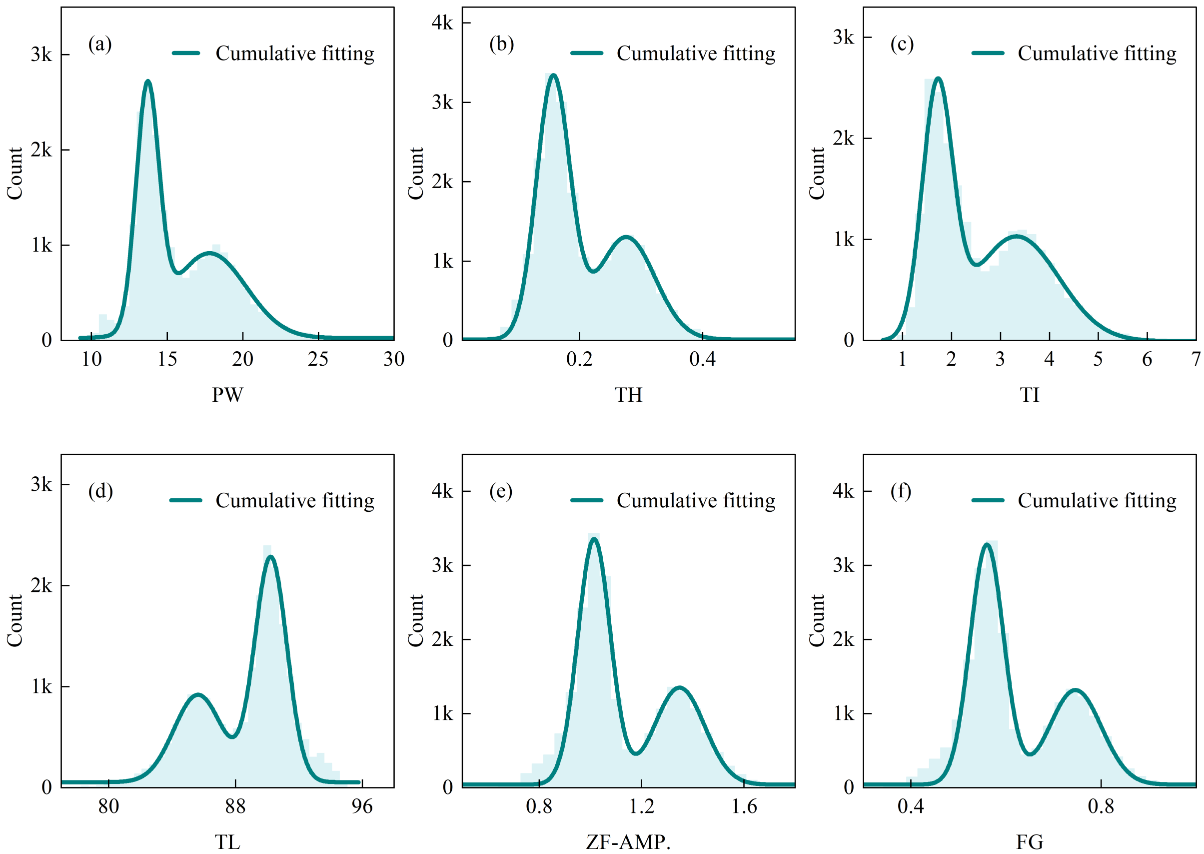

Fig. 10. The feature distribution of the low-energy domain.

feature data) in the low-energy domain dataset are selected as the training set, and the remaining 3,393 sets of feature data are used as the test set 1 for further discrimination; a total of 16,931 sets of feature data (including 8,564 sets of neutron feature data, 8,367 sets of γ-ray feature data, and 1,080 sets of γ-ray feature data) are selected from the high-energy domain dataset, and the remaining 1605 sets of feature data are selected as the test set 2 for further discrimination. In the subsequent discrimination experiments, we will focus on the residual test sets to compare the discrimination effect of the model on them.

C. CCM and LDA

In this study, two methods, the classical charge comparison method (CCM) as well as the LDA algorithm, are chosen for the experiments in order to compare them with KNN-LDA.

CCM uses the integral ratio of different intervals of the pulse signal to discriminate the particle type. Since the CCM more fully utilizes the shape difference between neutron and γ-ray pulses, it has a good discrimination effect under most conditions, and therefore it is also widely used by researchers. In this paper, the ratio of the short-gate integral Q_S to the long-gate integral Q_L is utilized as the discrimination factor of CCM, as in Eq. (17).

$$PSD_{CCM} = \frac{Q_S}{Q_L} \tag{17}$$

Firstly, we validate the training set data using CCM, as shown in Fig. 15, the CCM histograms of the two training sets, in which the FOM in the low-energy domain reaches 1.20 and the FOM in the high-energy domain reaches 1.15, which are both with good separation effect. Using the trough as the discrimination threshold, the classification results of the two datasets are fully consistent with the labeling of the feature criterion. The above illustrates that the training set data instantly divided by the feature criterion is very reliable and can be used for model training.

Then, the two test sets were subjected to CCM discrimination, and the resulting distribution histograms are shown in Fig. 16. As can be seen from the figure, the bimodal separation obtained from the test set in the low-energy domain is very poor, with a FOM of only 0.58; while the high-energy domain is relatively better, but the FOM is also only 0.89. Obviously, it is very difficult to achieve satisfactory discrimination results by CCM for the impulsive data which are difficult to differentiate, especially in the low-energy domain.

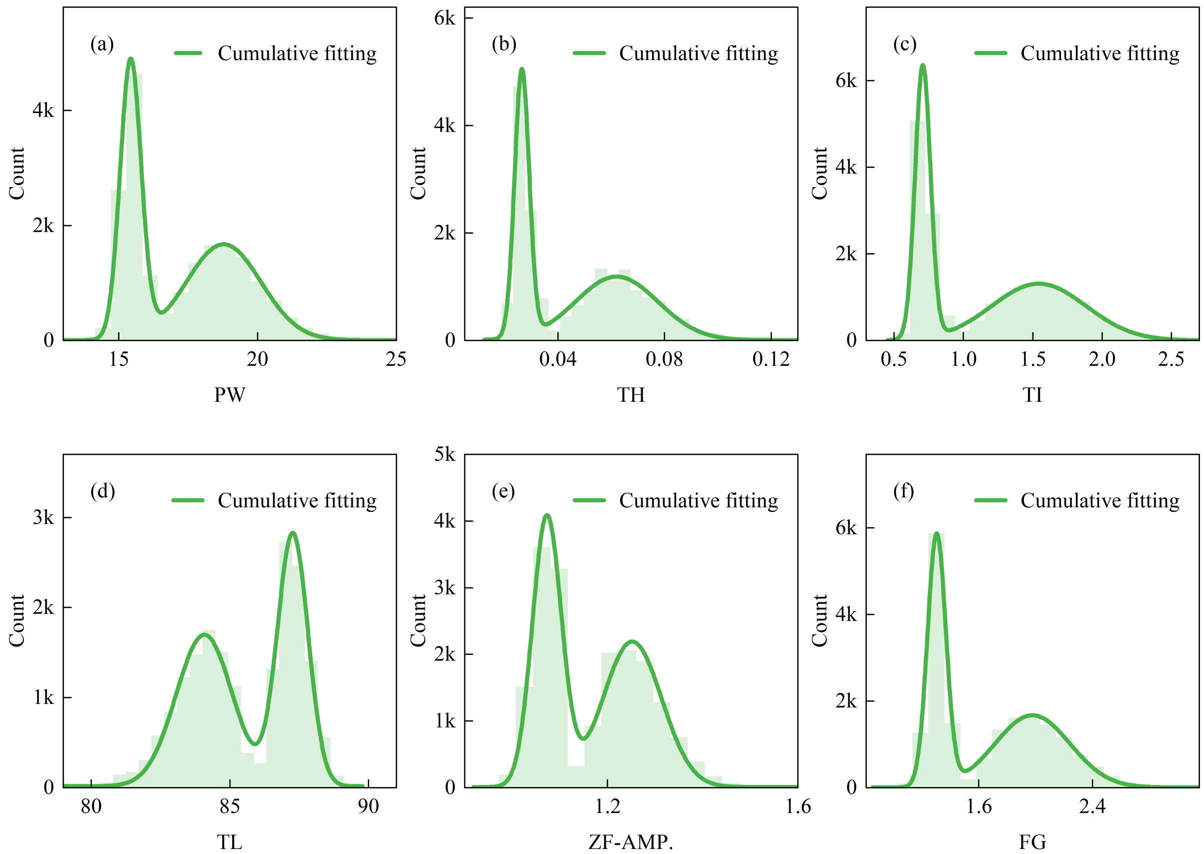

Fig. 11. The feature distribution of the high-energy domain.

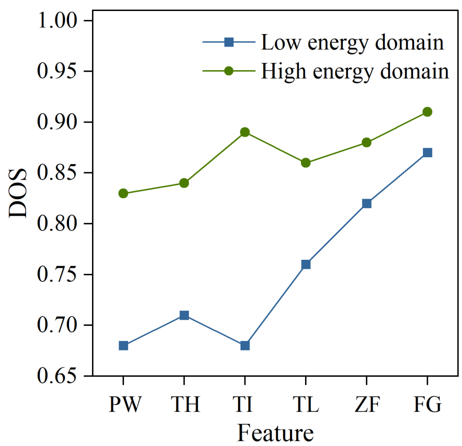

Fig. 12. Line chart of DOS.

Finally, the training set feature data is input into the LDA model for training, and the test set feature data is reduced to one-dimensional feature data according to the best projection direction obtained, and the resulting distribution histogram is shown in Fig. 17. As can be seen from the figure, the discrimination effect of the LDA model is better than that of the CCM, but the FOM in the low-energy domain is only improved by 0.08 to 0.65, while in the high-energy domain the FOM is improved by 0.23 to 1.12. The improvement of the discrimination effect of the LDA model is more obvious for the part of the high-energy domain, but it is still insufficient for the part of the low-energy domain.

D. KNN-LDA

The KNN-LDA algorithm needs to set a suitable value of nearest neighbor K. The value of K has a large impact on the performance and computational cost of the algorithm. Considering the computational efficiency of the algorithm as well as the feature dataset with high-density distribution characteristics, setting a smaller value of K satisfies the needs of the algorithm. The range of K value is set to [3,5,7], and the training set with labels is used to verify that KNN has the best classification accuracy when K=5. Therefore, in this paper, the K value is set to 5, that is, the mean value of the five training samples closest to the input sample is used as the regression of that input sample in KNN regression. After that, the training set data is input into the KNN-LDA model for train-

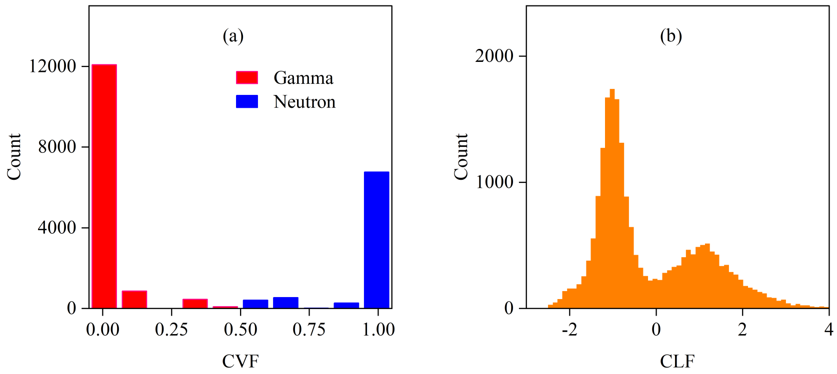

Fig. 13. Histogram of CVF and CLF of the low-energy domain.

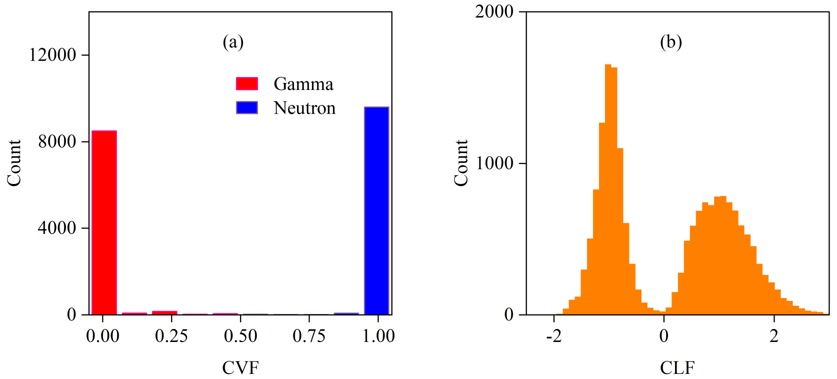

Fig. 14. Histogram of CVF and CLF of the high-energy domain.

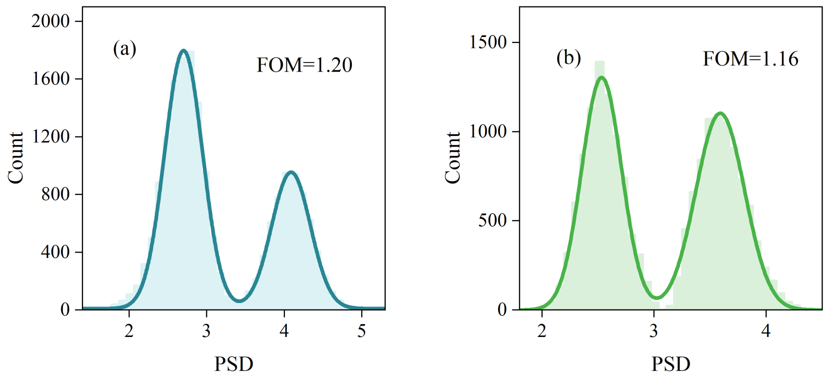

Fig. 15. Histogram of CCM of training set: (a) Low energy domain; (b) High energy domain.

ing, after which the test set data is regressed and downscaled, and the one-dimensional feature distribution histogram output from the model is shown in Fig. 18.

As seen in Fig. 18, the discrimination effect of the KNN-LDA model is excellent, and the double peaks are clearly distinguished in both low-energy and high-energy domains. And the FOM reaches 2.64 in the low-energy domain and 3.07 in the high-energy domain, which is a significant improvement of FOM.

E. Results and analysis

The FOM indicators obtained from each method were summarized as shown in Table 2.

As can be seen from Table 2, KNN-LDA has a brilliant

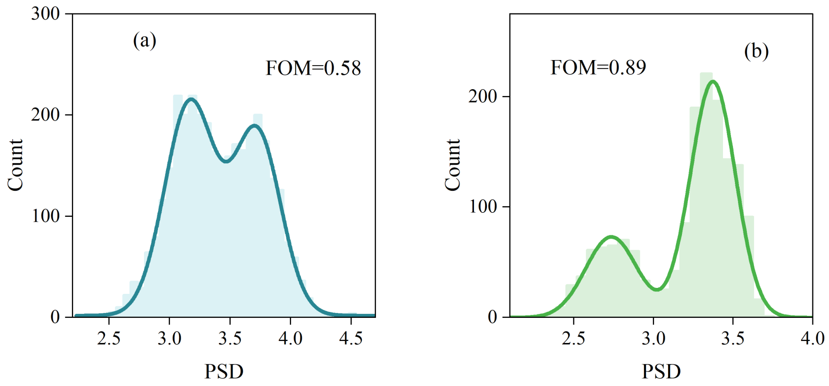

Fig. 16. Histogram of CCM of test set: (a) Low energy domain; (b) High energy domain.

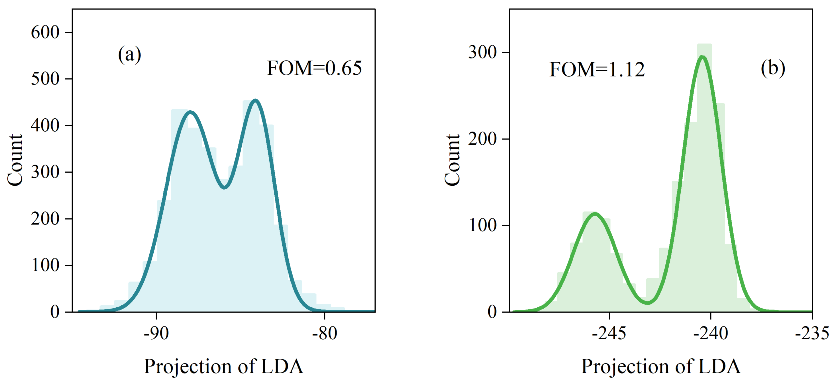

Fig. 17. Histogram of LDA of test set: (a) Low energy domain; (b) High energy domain.

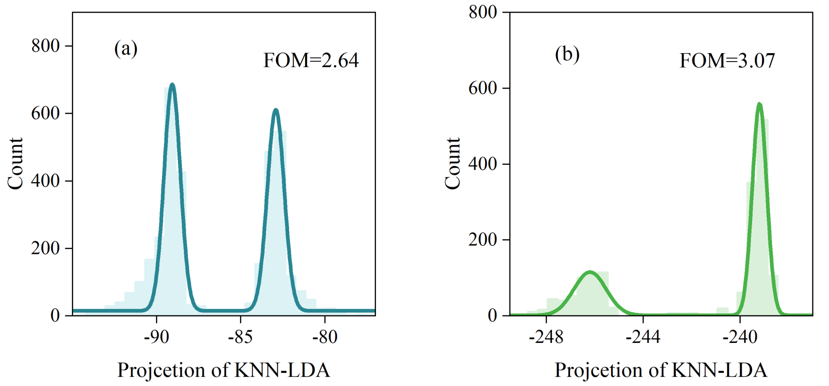

Fig. 18. Histogram of KNN-LDA of test set: (a) Low energy domain; (b) High energy domain.

FOM performance, in the test set of high-energy domain, the FOM obtained from the KNN-LDA model is 2.18 higher than the CCM, an enhancement of about 245%, and 1.95 higher than the LDA, an enhancement of about 174%; in the test set of low-energy domain, the FOM obtained from the KNN-LDA model is 2.06 higher than the CCM, an enhancement of about 355%, and 1.99 higher than the LDA, an enhancement of about 306%. It can be seen that KNN regression can effectively optimize the feature data in the multidimensional feature space, which in turn allows the KNN-LDA model to significantly improve the discrimination effect, especially for the low-energy domain part of the enhancement is very powerful. In addition, comparing the FOM value of different energy domains of the same method, the high-energy domains are all significantly better than the low-energy domains, which again verifies that the high-energy domain pulse data have more

Table 2. Experimental results.

Method	Energy domain	FOM
CCM		0.58
LDA	Low energy domain	0.65
KNN-LDA		2.64
CCM		0.89
LDA	High energy domain	1.12
KNN-LDA		3.07

obvious waveform differences and are easier to be discriminated, even though they all belong to the part that is more difficult to be discriminated in the original dataset.

The experimental results can be analyzed as follows: First, the proposed multi-features-based KNN-LDA model shows better discrimination performance compared to the traditional method based on a single feature, especially for low-energy-domain data, which demonstrates the advantage of multi-features comprehensiveness and supports the reasonableness of the KNN-LDA model. Second, the KNN-LDA model effectively mines and optimizes the discrimination ability of multi-features. And the strategy of constructing a training set for the supervised model from the multi-features in the original dataset and then performing high-precision classification using the model is both effective and feasible. This not only facilitates the application of the supervised model in field discrimination but also makes algorithm transplantation feasible to some extent. Lastly, the automatic feature extraction algorithm designed in this paper can effectively obtain the optimal feature distribution, which not only optimizes the feature extraction process but also greatly enhances the feasibility of this method.

IV. CONCLUSION

This paper proposes an n/γ discrimination model based on multi-features and KNN-LDA. The model achieves integrated automation, where feature extraction, dataset splitting, model training, and test set discrimination are performed continuously and automatically, offering a complete system with diverse functionalities. The experimental results show that the model achieves a FOM of 3.07 on the high-energy domain test dataset and 2.64 on the low-energy domain test dataset, significantly improving discrimination performance. The model effectively discriminates low-energy domain pulses that are difficult for CCM to distinguish, which provides a new idea for the introduction of supervised learning algorithms in the field of n/γ discrimination to solve the problem of the difficulty of discrimination in the low-energy domain, and also makes the transplantation application of supervised discrimination model more possible.

[1] D.Z. Ding, C.T. Ye, Z.X. Zhao, *et al.*, in *Neutron Physics: Principles, Methods, and Applications*, 2nd edn. (Atomic Energy Press, Beijing, 2005), p. 105

[2] A. Yamazaki, K. Watanabe, A. Uritani, *et al.*, Neutron-gamma discrimination based on pulse shape discrimination in a Ce:LiCaAlF$_6$ scintillator. NUCL. INSTRUM. METH. A. **652**, 635-438(2011). DOI: 10.1016/j.nima.2011.02.064

[3] J.X. Li, H.L. Hou, Y.F. Huang, *et al.*, Pulse-shaping method for real-time neutron/gamma discrimination at low sampling rates. Nucl. Sci. Tech. **34**, 165(2023). DOI: 10.1007/s41365-023-01306-z

[4] Bo. W, Gang. L, Kun. L, *et al.*, Digital pulse shape discrimination technique for liquid scintillation detector EJ-301. JINST. **14**, T11001(2019). DOI: 10.1088/1748-0221/14/11/T11001

[5] J. Choi, J. Park, J. Son, *et al.*, Study of n/γ discrimination using ^{3}He proportional chamber in high gamma-ray fields. NUCL. ENG. TECHNOL. **51**, 263-268(2019). DOI: 10.1016/j.net.2018.08.013

[6] D. Zhao, S. Feng, C. Hu, *et al.*, Characterization of the neutron/γ-ray discrimination performance in an EJ-301 liquid scintillator for application to prompt fission neutron spectrum measurements at CSNS. RADIAT. MEAS. **151**, 106703(2022). DOI: 10.1016/j.radmeas.2022.106703

[7] F. Wang, M. Yang, J. Wang, *et al.*, A comparison of small-batch clustering and charge-comparison methods for n/γ discrimination using a liquid scintillation detector. RADIAT. MEAS. **1028**, 166379(2022). DOI: 10.1016/j.nima.2022.166379

[8] L. Liu, H. Shao, Study on neutron–gamma discrimination method based on the KPCA-GMM. NUCL. INSTRUM. METH. A. **1065**, 168604(2023). DOI: 10.1016/j.nima.2023.168604

[9] W. Hu, G. Zhang, Y. Zhang, *et al.*, Neutron/gamma (n/γ) discrimination method based on KPCA-MPA-ELM. Nucl. Tech. **47**, 75-84(2024). DOI:10.11889/j.0253-3219.2024.hjs.47.040403

[10] J. Yan, R. Liu, C. Li, *et al.*, A comparison of n/γ discrimination by the rise-time and zero-crossing methods. SCI. CHINA PHYS. MECH. **53**, 1453-1459(2010). DOI: 10.1007/s11433-010-4036-8

[11] T. Szczcsniak, M. Grodzicka, M. Moszynski, *et al.*, Digital Neutron-Gamma Discrimination Methods: Charge Comparison versus Zero-Crossing, in IEEE Nuclear Science Symposium and Medical Imaging Conference Record, pp.1-4(2014). DOI: 10.1109/NSSMIC.2014.7431222

[12] M. Nakhostin, A comparison of digital zero-crossing and charge-comparison methods for neutron/γ-ray discrimination with liquid scintillation detectors. NUCL. INSTRUM. METH. A. **797**, 77-82(2015). DOI: 10.1016/j.nima.2015.06.041

[13] T. Ding, Y. Jiang, X.Wang, *et al.* , Study on neutron-gamma discrimination methods based on GMM-KNN and LabVIEW implementation. Nucl. Sci. Tech. **35**, 84-99(2024). DOI: 10.1007/s41365-024-01545-8

[14] S. Yoon, C. Lee, H. Seo, *et al.*, Improved fast neutron detection using CNN-based pulse shape discrimination. NUCL. ENG. TECHNOL. **55**, 3925-3934(2023). DOI: 10.1016/j.net.2023.07.007

[15] S. Zhang, Z.Wei, P. Zhang, *et al.*, Neutron-gamma discrimination with broaden the lower limit of energy threshold using BP neural network. APPL. RADIAT. ISOTOPES. **205**, 111179(2024). DOI: 10.1016/j.apradiso.2024.111179

[16] X. Yu, J. Zhu, S. Lin, *et al.* Neutron–gamma discrimination based on the support vector machine method.

NUCL. INSTRUM. METH. A. **777**, 80-84(2015). DOI: 10.1016/j.nima.2014.12.087

[17] H. Arahmane, A. Mahmoudi, E.-M. Hamzaoui, *et al.*, Neutron-gamma discrimination based on SVM combined on nonnegative matrix factorization and continuous wavlet transform. MEASUREMENT. **149**, 106958(2020). DOI: 10.1016/j.measurement.2019.106958

[18] H. Arahmane, E.M. Hamzaoui, Y. Ben Maissa, *et al.*, Neutron-gamma discrimination method based on blind source separation and machine learning. Nucl. Sci. Tech. **32**, 18(2021). DOI: 10.1007/s41365-021-00850-w

[19] S. Yousefi, L. Lucchese, M.D. Aspinall, Digital discrimination of neutrons and gamma-rays in liquid scintillators using wavelets. NUCL. INSTRUM. METH. A. **598**, 549-553(2009). DOI: 10.1016/j.nima.2008.09.028

[20] K.N. Li, X.P. Zhang, Y. Li, *et al.*, Investigation on n/γ Discrimination Methods for Liquid Scintillator Detector. At. Energy Sci. Technol. **5**, 913-919(2014). DOI: 10.7538/yzk.2014.48.05.0913

[21] R.C. Pereira, A. Fernandes, N. Cruz, *et al.*, Neutron/Gamma discrimination code based on trapezoidal filter. FUSION. ENG. DES. **134**, 118-122(2018). DOI: 10.1016/j.fusengdes.2018.07.002

[22] G.F. Liu, M.J. Joyce, X.D. Ma, *et al.*, A Digital Method for the Discrimination of Neutrons and γ Rays With Organic Scintillation Detectors Using Frequency Gradient Analysis. IEEE T. NUCL. SCI. **57**, 1682-1691(2010). DOI: 10.1109/TNS.2010.2044246

[23] G. Liu, J. Yang, X.L. Luo, *et al.*, A comparison of different discrimination parameters for the DFT-based PSD method in fast scintillators. RADIAT. MEAS. **58**, 12-17(2013). DOI: 10.1016/j.radmeas.2013.07.008

[24] B. D'Mellow, M.D. Aspinall, R.O. Mackin, *et al.*, Digital discrimination of neutrons and γ-rays in liquid scintillators using pulse gradient analysis. NUCL. INSTRUM. METH. A. **578**, 191-197(2007). DOI: 10.1016/j.nima.2007.04.174

[25] J. Cai, D. Li, P. Wang, *et al.*, Fast pulse sampling module for real-time neutron–gamma discrimination. Nucl. Sci. Tech. **30**, 84(2019). DOI: 10.1007/s41365-019-0595-1

[26] S. Peng, Y. Yi, Y. Lu, *et al.*, Neutron/Gamma Real-time Discrimination Based on Frequency Comparison Analysis. At. Energy Sci. Technol. **30**, 520-525(2020). DOI: 10.7538/yzk.2018.youxian.0622

[27] S. Liu, W. Zhang, Z. Zhang, *et al.*, Performance of real-time neutron/gamma discrimination methods. Nucl. Sci. Tech. **34**, 8(2023). DOI: 10.1007/s41365-022-01160-5

[28] L.F. Liu, Research on n/γ discrimination method of EJ309 liquid scintillator detector. Master Dissertation. Beijing:Academy of Military Sciences, 2022. DOI:10.27193/d.cnki.gjsky.2022.000054

[29] P.V. Chuan, L.V. Tung, N.N. Hai, *et al.*, A Study on the Impact of Pulse Shaping Parameters on Zero-Crossing Method Performance for Neutron/Gamma Discrimination. IEEE T. NUCL. SCI. **70**, 2464-2470(2023). DOI: 10.1109/TNS.2023.3323369

[30] Y. Jiang, T. Ding, X. Jiang, Automatic n-γ discrimination design based on feature intelligent extraction. Nuclear Electronics and Detection Technology. **44**, 1091-1100(2024). DOI:10.20173/j.cnki.ned.20240903.001

[31] W. Bo, L. Gang, L. Kun, *et al.*, Digital pulse shape discrimination technique for liquid scintillation detector EJ-301. J. INSTRUM. **14**, T11001(2019). DOI:10.1088/1748-0221/14/11/T11001

[32] X. Wang, Q. Tang, X. Jiang, Simulation of n-γ pulse signal discrimination based on KNN classification algorithm. Electronic Measurement Technology. **45**, 164-170(2022). DOI: 10.19651/j.cnki.emt.2209025

[33] K. Li, X. Zhang, M. Zhang, *et al.*, n/γ Pulse Shape Discrimination in Cs2LiYCl6:Ce3+ Crystal Using Fisher Linear Discriminant. At. Energy Sci. Technol. **51**, 2069-2074(2017). DOI: 10.7538/yzk.2017.51.11.2069

[34] Z.H. Zhou, in *Machine learning eng*, ed. by H. Xue (Peking University Press, Beijing, 2016), pp. 60-63

[35] M.E. Iranian, M. Mohseni, S. Aghili, *et al.*, Real-Time FPGA-Based HIL Emulator of Power Electronics Controllers Using NI PXI for DFIG Studies. IEEE J. EM. SEL. TOP. P. **10**, 2005-2019(2022). DOI: 10.1109/JESTPE.2020.3023100

[36] G. Dhanabalan, S.T. Selvi, Software Design of VerilogHDL Code Generation for Ladder Diagram and Data Acquisition Using LABVIEW. WIRELESS PERS. COMMUN. **128**, 1087-1115(2023). DOI: 10.1007/s11277-022-09990-7

[37] R. Swarnalatha, S. Athiban, Analysis of brain wave data to detect epileptic activity using LabVIEW. SOFT COMPUT. **27**, 17231–17241(2023). DOI: 10.1007/s00500-023-08047-6